DOCUMENT RESUME

ED 129 908 95 TM 005 751

AUTHOR Follettie, Joseph F.
TITLE GPO: Send Me The Primary Effects of Common Instruction! Professional Paper 34.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO SWRL-PP-34
PUB DATE 10 Mar 76
CONTRACT NE-C-00-3-0064
NOTE 46p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; City Wide Programs; Criterion Referenced Tests; *Educational Accountability; *Educational Assessment; Elementary Secondary Education; Instructional Programs; *National Programs; Norm Referenced Tests; Program Effectiveness; Standardized Tests; State Programs; Taxonomy; Testing Problems; *Testing Programs
IDENTIFIERS Instructional Hierarchies; National Assessment of Educational Progress

ABSTRACT

General features of local and national programs for assessing achievements referencing the common instruction are discussed within a single mastery achievement testing framework. The envisioned programs differ only in informative detail. Most such differences are viewed as amenable to formalization and the basis for distinguishing between local instructional management requirements and state and national stocktaking requirements for information on scholastic achievements is illustrated for selected knowledges and skills. The implications of the envisioned achievement testing programs for local, state, and national determinations of educational productivity are noted. It is contended that the earliest apt educational productivity estimates must be based on aggregate direct costs of education as inputs--perhaps with a "catch-up" cost portion removed by general agreement--and short-term absolute scholastic achievement effects as outputs. (Author)

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

"GPO: Send Me The Primary Effects of Common Instruction!"

10- March 1976 | Professional Paper 34

The work upon which this document is based was performed pursuant to Contract NE-0-00-3-0064 with the National Institute of Education. SWRL reports do not necessarily reflect the opinions or policies of the sponsors of SWRL R&D.

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

Professional Paper 34

10 March 1976

"GPO: SEND ME THE PRIMARY EFFECTS OF COMMON INSTRUCTION!"

Joseph F. Follettie

ABSTRACT

General features of local and national programs for assessing achievements referencing the common instruction are discussed within a single mastery achievement testing framework. The envisioned programs differ only in informative detail. Most such differences are viewed as amenable to formalization and the basis for distinguishing between local instructional management requirements and state and national stocktaking requirements for information on scholastic achievements is illustrated for selected knowledges and skills. The implications of the envisioned achievement testing programs for local, state, and national determinations of educational productivity are noted. It is contended that the earliest apt educational productivity estimates must be based on aggregate direct costs of education as inputs--perhaps with a "catch-up" cost portion removed by general agreement--and short-term absolute scholastic achievement effects as outputs.

Contents

"GPO: SEND ME THE PRIMARY EFFECTS OF COMMON INSTRUCTION!"

Joseph F. Follettie

The Government Printing Office's reply to this request would have to be, "Please inquire again in 1985. Based on extrapolation from trend, this is the earliest expected date for the document. If earlier delivery is required, appropriate other officials should be contacted immediately."

Educational accountability is not yet tenable because output is not usefully defined and assessed or estimated. This is so in spite of the fact that much ink has flowed on how scholastic achievements should be assessed. This literature for the most part implicates pterodactyloid assessment, institutionalized in levitating, reptilian, widely-used, standardized, norm-referenced, national achievement tests (NRTs). Many millions of student hours per year are spent on negotiating NRTs. Yet the large expenditures of time and money devoted to this enterprise yield no data pertinent to rendering an account of the primary effects of common instruction.

NRTs fail as instruments for gauging achievement resources because they define knowledge-skill domains too generally, use item-selection practices that relativize proficiency domains, and use test-norming practices that relativize test scores with respect to time. Consider these defects in the context of word-meaning proficiency assessment. Let items take the ancient and honored vocabulary test-item form, "definiens + response alternatives."

Every definiens falls on scales for morphological and semantic complexity. NRT developers need not--and owing to item selection practices grounded on the doctrines of intermediate difficulty and high discrimination indeed cannot--say what level or range of morphological-semantic complexity the set of used definiens reflects. The used items simply are those which are of intermediate difficulty for a given age level in a given space-time era. A student can be proficient or not with respect to featured items, but proficiency along scales for complexity cannot be determined because items are not selected nor tests constructed to enable such determinations.

In defense of this arrangement, NRT advocates argue in non sequitur fashion that proficiency assessment must be so relativized because none can know the totality of proficiencies which lurk in a skull. Yet, given a suitable prior test construction effort, one can of course determine that a respondent responds appropriately to most words like "dog" and "hit," responds appropriately to some words like "laser" and "theory," and bombs out when responding to words like "cathexis" and "campanology."

Response alternatives might be pictorial illustrations, diagrammatic representations, potential definienda, sentences using the definiens, or a combination that jointly tests formal and semantic understanding. While these and other response forms class under a broad vocabulary-skills rubric, outcomes of testing might vary with the response form used. NRT advocates typically classify such distinctions as nit--sometimes correctly so, but largely on the basis of factor-analytic tools that place elephants and grains of sand under the same heading. The fuzziness of NRT proficiency domains due to liturgical item-selection and test-norming practices is enhanced by the freedom test developers allow themselves during item-form selection. A sampling approach to assessing achievements for state and national stocktaking purposes simply sanctions wider choices concerning presentation and response forms.

An alternative to NRTs is the National Assessment of Educational Progress (NAEP)--discussed in a later section. Data thus far collected by NAEP are meager, but the project has had only a brief history. The larger problem facing NAEP is that is has not yet appreciably managed to shed an incestuous bondage to norm-referenced testing rationale. We cannot expect the presses soon to roll for The Primary Effects of Common Instruction because the prevailing NRTs--e.g., Comprehensive Tests for Basic Skills, Metropolitan Achievement Tests, Stanford Achievement Tests--and NAEP as presently formulated are inconsonant with developing a data base useful to educational policymaking. The absence of suitable tests is not, however, the only barrier to the development of an adequate data base for this purpose. The formulation and execution of a suitable program also is hampered by concerns expressed under the rubrics of local control over instruction and confidentiality of records.

Local control over school offerings is grounded in national history and has constitutional foundations. Yet it is noteworthy that so little substate local control of instruction presently exists. Much instruction is mandated at a state level. These mandates in statutory form tend to be general. The interaction between state bodies, textbook publishers, and organized segments of society render instructional mandates more specific--cf, Broudy (1975), whose assertions concerning forces molding textbooks probably apply equally to prevailing national achievement tests. Local control is at once a worthwhile objective and a slogan used to mask the thwarting of the general interest in an education which tries to underwrite effective personal functioning.

If the nation's boundaries have more cultural significance than those imposed about the world by the colonial powers during the 19th Century, our society is characterized by some common denominator for knowledges and skills going beyond apple pie and celebrations of greatness, liberty, charity, and justice. Even though a consensus of special interests heretofore has been allowed to define the common

denominator and to install this special view as nationwide instruction, it is apparent that much of the nation's common instruction--e.g., in reading and other communication skills, in mathematics--serves the general interest. Where local schools are allowed to augment state mandates on instruction, in recent years they have tended to introduce some part of life as it is lived into their offerings--witness the growing popularity of consumer economics. The specification of common knowledges and skills which should be assessed nationwide does not threaten the existing miniscule local control over education. It only threatens those who use the local-control slogan to screen the exercise of power.

Federal mandates are no less sensitive to special interests than are state mandates. The competition for the political and economic control of education at the various levels of government need not, however, thwart concurrent professional efforts to better address common instruction and to improve the effects of this large-scale public enterprise.

Some worry that effective estimation of national achievement resources risks an invasion of privacy culminating on production of a dossier for each child in school. Such inquisitiveness can be proscribed by law. Moreover, economic considerations and sampling technology coalesce to insure that data-source confidentiality will not be breached. Past, present, and projected budgets for the National Center for Educational Statistics suggest all studies which it might fund save those for enrollments, property count items, and direct costs must entail sampling. As the national polls demonstrate, sampling studies often breach ignorance using few respondents whose identities are of interest only for technical follow-up purposes.

Intermeshed with the objective of apt stocktaking for scholastic achievements at state and national levels is the issue concerning accountability of the large public component of the education enterprise. The year-to-year productivity of education cannot be effectively estimated until output is definitively modelled on a periodic basis. There is room for disagreement concerning how educational productivity should be defined. But opposing, in principle, the concept of evaluating education on productivity grounds is increasingly difficult to defend. Fortunately for foot-draggers, the concept can be effectively opposed in practice, since the only "output" measure yet widely available is student years of schooling.

A useful national stocktaking program should reflect accomplishments of the schools and should provide information which educators can use for a better purpose than getting on the right side of a zero-sum game. The policymaker's need for definitive information on the outcomes of common instruction is no greater than the classroom teacher's need for readily-obtained, cost-effective information on student progress. Frequent proficiency assessment in the classroom provides teachers with a basis for fine-tuning the management of instruction.

This paper posits that periodic stocktaking to serve policymaking objectives and more frequent testing to serve classroom instructional management objectives can be met within a single conceptual framework. An achievement testing perspective is sketched which serves both functions while assuring local control over the details of instruction and providing a local basis other than age/grade for matching students to national tests.

The first section of the paper addresses a variety of background issues underlying specification of educational productivity. The second section sketches the possible origins of spiraling costs of education and concludes that future such increases in part will need be justified on the basis of demonstrably-increasing output. The third section discusses the NAEP program and factors militating against the GPO response, "Dear Citizen: Enclosed NAEP findings respond to your request." Remaining sections sketch an achievement testing perspective serving national and local needs consonant with educational productivity considerations.

EFFECTIVENESS AND PRODUCTIVITY ISSUES

Current Data Base

While most public enterprises appear headed toward productivity accounting, neither input-costing nor output-specifying practices yet permit this. For education, the input-costing problems are less vexing than the problems of specifying and quantifying outputs but remain challenges to be overcome. The aggregate direct costs of education--operating expenditures, interest, plant expansion outlays --presently are reported and probably adequately so--cf, National Center for Educational Statistics (1975, Chapter 3). Breakdowns of direct costs by grade level, instructional program, or type of output are not routinely reported. Occasional special studies--cf, Thomas (1971, Chapter 5)--provide such information for a given locale, school year, and facet of education. The indirect costs of education--e.g., foregone taxes and student earnings--also are not routinely reported. Occasional special studies--cf, Thomas (1971, Chapter 3)--provide estimates of indirect costs for a given state and school year. Existing aggregate direct-cost data might suffice for earlier attempts to gauge educational productivity. In time, new cost-accounting practices which more analytically relate costs to programs probably will be required.

The achievement return on school expenditures hangs somewhere between fragmentary--cf, NCES (1975, Chapter 2)--and conjectural, in spite of more than a century of efforts to judge education from a productivity perspective--cf, Wynne (1972, Chapter 3). The widely-used norm-referenced national achievement tests provide the conjectural basis. NAEP findings provide the fragmentary basis.

Education researchers--a host of investigators whose diverse points of view are reflected by such writers as Gagné (1970), Popham (1971), Wynne (1972), and Vandermyn (1974)--specify or advocate particular views of achievement that might be apt to specification of output. While germane to the quest for better output models, such notions as the operationalizing of proficiencies through specification of ordered instructional hierarchies and the assessment of proficiencies so specified using domain- and criterion-referenced tests alone do not guarantee that pertinent outputs will be apprehended and assessed.

Productivity Functions

Educational productivity can be alternatively defined. Education economists--e.g., Benson (1961), Rogers and Ruchlin (1971), Thomas (1971)--have proposed a variety of entertainable education productivity functions. The weakest of these is a cost-effectiveness function, which typically defines output on educational time in grade. If Program A uses one student year of schooling and Program B uses a fraction of a year, Program B is the more cost-effective. Cost-effectiveness can reasonably be used to evaluate noninstructional options--e.g., in the equipment domain. Where achievement is insufficiently well-defined, doing nothing instructionally tends to be optimally cost-effective.

An input-output function defines output on a primary (scholastic) effect of instruction--e.g., word-sounding skill referencing a scale of word complexity. As used by regression analysts in studies of antecedents of primary effects of education, this function typically does not reflect costs as inputs. Rather, the antecedents are school and background factors. Such a function in principle is useful to optimizing educational output.

A cost-benefit function defines output on a secondary economic effect of instruction. Earnings are studied as a function of prior schooling; additional earnings are studied as a function of additional prior schooling. In such studies, the year of schooling is a cost input rather than the pseudo-output it is in a typical cost-effectiveness study. Such a function references one of the secondary effects of possible interest but does not touch educational productivity defined on primary effects. An alternative function of this class--e.g., one wherein benefits are defined on adult coping behaviors--probably requires definition.

A function not yet much considered by economists takes costs from the cost-benefit function and output from the input-output function, yielding a cost-output function, a primary-effects analog to the cost-benefit function. The immediate problem posed by a current cost-benefit analysis is that findings indicate the productivity of education occurring several years ago. I don't believe

that a basis will exist for predicting future cost-benefit for present education until the relation between current cost-output and past cost-benefit data is established. Such a basis, if feasible, entails several years of cost-output findings.

When considering the current education enterprise, early attempts to determine educational productivity probably must feature a cost-output function predicated on primary-effects assessment. The cost information underlying cost-output estimation for education probably must increase in specificity. The greater and more immediate challenge apparently is on the output side, where achievement information must increase in specificity by gravitating to a domain-referenced testing level whose present meaning is discussed in later sections.

While some view the issue of educational productivity primarily as one of overcoming public apathy and power-conservation tendencies of school officials at all levels, a more serious immediate problem is that the necessary tools must yet be crafted or refined. Many promising beginnings are on hand. Yet it is a myth propounded by zealots operating well ahead of the leading edge of the pertinent states-of-the-art that the necessary tools are on the shelf, ready for use when a propitious instant arrives. Such an instant might well have been forced at any time during the last decade had the input and output armamentaria been up to the waging of a determined productivity campaign.

Trend in Social Stocktaking

Economic policymaking is grounded on an economic report whose indicators enable detection of changes in economic conditions. There is as yet no comparable social report. The social indicators now in use are fragmentary and implicate data open to alternative interpretations or to no rational interpretation whatsoever. Policymakers are unsure of the state of the union for a variety of social conditions to which legislation responds. The data gap is particularly acute for education. While serving polemicists of all political persuasions, the failure yet to come to grips with educational outputs defeats the wise use of resources.

Increasing tendencies on the part of the public to ask public enterprises to justify their claims on revenues render it probable that an accountability system lurks in education's future. The remaining debates center on how to make such a system fair and when the system can be installed. The education research community now must come to terms with the eventuality that an underfair system will be installed if the quest for distinctive assessment toys does not give way to a quest for apt proficiency assessment tools.

The philosophic-methodological issues underlying social reporting have received systematic study only during the last few decades. Bauer's Social Indicators (1966)--an edited work growing out of attempts to gauge the impact of NASA programs on national life--illustrates the conceptual problems. Bauer and his associates distinguish between primary effects--essentially the immediate consequences of specified programs--and secondary effects--deferred consequences to which specified and other programs contribute. The secondary effects of any program might well run the gamut for social reporting. For education at least, even the antecedents of primary effects are in dispute.

The U. S. Department of Health, Education, and Welfare's Toward a Social Report (1969) is a first programmatic sketch of a comprehensive system of social indicators. Social indicators legislation--cf, Wynne (1972, Chapter 8)--is wending its way through Congress at an inexorable but disjointed incremental rate. Against this background, the 1969 DHEW report notes that

> The Digest of Educational Statistics, for example, contains over a hundred pages of educational statistics in each annual issue, yet has virtually no information on how much children have learned (p. 66).

Whatever our hopes for the NAEP program, this conclusion is very nearly as warranted today as it was when made--cf, NCES (1975, pp. 22-30, 137-143).

The DHEW report cites fragmentary evidence supporting the view that scholastic achievement has increased in recent decades.

> The Educational Testing Service recently assembled 186 instances in which comparable tests have been given to large and roughly representative national samples of students at two different times over the past two decades. In all but 10 of the 186 paired comparisons, the latter group performed better than the earlier group. On the average an additional eight percent of the students in the more recent group scored higher than the median student in the old group (p. 67).

While such fragments are comforting, the evidence must be balanced against the fact that real per capita costs of education more than doubled during the period from 1950 to 1970. Some of this increase in constant dollars reflects catch-up costs for teachers insufficiently paid in 1950. The possible origins of increased costs are discussed in the next section.

Primary Effects

The absence of an adequate definition of primary and secondary effects of education has not dissuaded investigation of the school's contribution to scholastic and deferred--primarily economic--achievements. Some investigators--e.g., Coleman (1966), Comber and Keeves (1973)--conclude that the school's contribution to production of primary educational effects is minor; others--e.g., Bowles and Levin (1968) referencing Coleman (1966), Coleman (1975) referencing Comber and Keeves (1973)--use the same data and find that the school's contribution is less minor, particularly outside the reading domain (Coleman, 1975). The impression one forms when reviewing such findings is that they are functions of rather simplistic views--whether imposed by study designs or the perspectives of regression analysts--concerning who critically influences whom. When studies grow more sophisticated for such matters, larger primary-effects contributions by the schools should be apprehended--see Hanson and Schutz (1975), the first of a planned series of reports. When schools become more sophisticated concerning their potential for usefully influencing parents, another increment is to be expected.

Studies that ask whether the school's contribution to primary effects are consequential must remain inconclusive until lines of influence are more clearly drawn and achievements are less vacuously and more inclusively specified.

The primary effects of education are achievement outputs--e.g., for word sounding and calculating and for using the telephone to effect transactions. Primary effects referencing such skills and knowledges in theory are and in practice usually are jointly determined by inputs occurring in and out of the classroom. Moreover, it is not precluded that the school will extend its influence into the home either to bring parents into a fruitful instructional partnership or to increase the personal accountability of students who are deliberately thwarting their own instructional accomplishments. This is a point not yet sufficiently acknowledged by input-output regression analysts --e.g., Comber and Keeves (1973), Coleman (1975). The aggregate-inputs models used by such investigators typically preclude "school variables" acting on "background variables." This is easy enough to justify when the background variable rides on a gene. But the home is more than a biological grouping. Typically it reflects parental authority. Its parents typically are interested in the scholastic welfare of their children and amenable to the proposition that teachers can contribute to parental effectiveness in assisting the scholastic growth of their children. Such school-to-home inputs and their refinements based on home-school interaction are potentially profound foundations for improved educational productivity, fairly distributed.

Secondary Effects

Some investigators--e.g., Jencks et al. (1972)--assert that the paramount secondary effect--whether of education or of the American political system in general--should be to increase social mobility and find that the schools contribute little to such an effort. Others --e.g., Northcutt (1974; 1975, in press)--imply that effective functioning in the common undertakings of adult life should be the paramount secondary effect of education and find that many adults--one-fifth or more--lack certain of the knowledges and skills required to so function. The society has yet to make its wishes crystal-clear regarding the secondary-effects objectives for elementary and secondary schools.

Studies that ask whether a school's contribution to alternative secondary effects are consequential must first consider just which secondary effects the schools could hope to contribute to and then probably should abandon the vacuous instructional time-in-grade view of school inputs heretofore characterizing such studies.

Three possible second-order effects of elementary-secondary education are to increase the proportion of students going to college, to affect social mobility, and to increase the extent to which individuals can deal responsibly and self-reliantly with the common undertakings of adult life. Present comments assume that the third of these possibilities constitutes the most realistic view of the longer-term objectives of education at elementary and secondary school levels. This is no way implies curtailment of educational opportunity, and this paper specifically rejects Boudon's (1973) view that "differences in level of educational attainment according to social background" constitutes "inequality of educational opportunity." Such a view begs the answer to an empirical question. Worth considering, however, are the following propositions:

- The common instruction on which desired secondary effects of education rest is well less than half of the offerings of elementary and secondary schools.

- Minimal standards for outputs addressed by the common instruction should be promulgated and achieved.

Such a view assumes a floor for common proficiencies below which effective adult functioning is imperiled. Unlike secondary-effects models reflecting economic productivity in adulthood, such a view treats primary and secondary effects of education as essentially isomorphic except for context.

It is entirely decent to hope that education in time will contribute to the lowering of the correlation holding between the socioeconomic status (SES) array for one generation and that for

the next. If nobility of thought is at issue, then one should hope that all individuals in time will know all worth knowing and be able to do all worth doing. However, one can in fairness ask education only to assist every individual to reach the highest profile for proficiencies of which he is capable. It is an empirical question whether schools that are productive and unassailably democratic will change SES rankings from one generation to the next, given the near-universality of the common education.

Mandates on Output

It appears tenable that the schools currently are underproductive and that they can and should raise absolute achievements for all but the highest achievers. The highest achievers for a given educational era--e.g., those who read at three and a half years, write at five years, do calculations at six and a half years, and sail through K-6 instruction in an independent studies mode--apparently function at a threshold level that cannot be surpassed until dramatic breakthroughs for pertinent knowledge usher in a new educational era. During the current era, it is fair to ask the schools to narrow achievement variance by progressively raising means and reducing the bounds for achievement distributions.

Instruction is input; its effect is output. The problem with current mandates referencing the common instruction is that they too often mandate input but fail to mandate achievement outputs which such instruction should instill. The schools typically are asked to invest the input but not to guarantee the output. Were all K-12 offerings mandated, it might be possible to guarantee the associated outputs only by holding some students in school for several years beyond the normal high school graduation age. Since well under half of these offerings form a common mandate--a condition that promises to continue in force--it appears tenable to guarantee outputs for the common instruction in the context of 12-13 years of schooling. Minimal output standards referencing the common instruction could be met.

Such standards aspire to do more than remove the correlation between parental SES and student output. In its limited domain, a perspective for minimal output standards seeks to remove all differences in adult coping proficiencies if this is humanly possible. Outside this restricted domain, students might be expected to continue to vary for types and numbers of proficiencies acquired. For the core proficiencies bearing on effective functioning in adulthood, there would be equality of output--not in the restricted sense advocated by Boudon but in a universal sense that does not inquire into parental SES.

Envisioned mandates for output seek to insure that no losing cold hands are dealt to students or that all students exit secondary

school able to play life in ways which they and the society find useful. However much it might exercise the educational system, such an objective does credit to a nation adopting it. This view implicates different basic knowledges and skills than a mathematician concerned with the entry proficiencies of the college-bound or a socialist concerned with narrowing economic differentials by decree might elect.

Not all of the technical problems hindering evaluating education on a productivity basis are on the output side. The disaggregation of human input--Nollen (1975)--is perhaps a more formidable undertaking in education than it is in other cost-accounting settings. Disaggregation of costs in general is a problem. For example, if it is noted that the per capita direct cost of K-12 education in constant dollars was almost two and a half times as great in 1970 as in 1950, school personnel will respond--with justification--that they were overworked and underpaid in 1950 and that some of the increased cost is a "catch-up" cost, or a redistribution of personal income that better reflects the relative worth of teachers and school officials. There probably is a necessity for distinguishing between those direct costs that assertedly reference catch-up and those that do not. National reporting in time should rely less on trade publications--e.g., those of NEA--for such determinations. Likely antecedents for constant-dollar increases in the per capita cost of education are discussed in the next section.

ORIGINS OF SPIRALING COSTS OF K-12 EDUCATION

Most of the direct costs of education occur under an "expenditures and interest" heading in census data. If one accepts the implicit price deflator used to transform current-dollar into constant-dollar values, the per capita expenditures and interest for K-12 education, in 1958 constant dollars, increased with positive acceleration from $240 in 1950 to $335 in 1960 to $575 in 1970. The per capita reference throughout this section is to a member of the K-12 enrollment. The present figures are approximations derived from economic and education data presented by Bureau of the Census (1974). NCES (1975, Chapter 3) presents more-sophisticated analyses and projections for the present decade. These data indicate that per capita cost in constant dollars should rise by 22% during the period from 1971-72 to 1977-78.[1]

---

[1] On page 36, NCES refers to federal categorical-aid programs as "propgrams." This subtle introduction of a new word into the lexicon may help bridge the current gap between revenue sharing and mandated magic in forwarding the public interest. In Table 18 (page 140), the report uses the rubric "Petroleum education." This term probably refers to the extent to which parental education lubricates the scholastic chances of their offspring. It is in the nature of all writers to editorialize.

Several factors are at work in the spiraling costs of K-12 education. A crude analysis of the contributions of the different apparently pertinent factors follows.

Since the cited figures are in constant dollars, the spiraling costs cannot be due to inflation. Since they are per capita referenced and exclude outlays for plant expansion, it is doubtful that increased capital outlays to accommodate increasing enrollment much influence the cost trend.

One factor in recent increasing costs of K-12 education is increasing federal expenditures, in such forms as ESEA Title I. Much such support has gone to programs attempting to raise the achievement of certain low-achieving components of the enrollment. The effects of such funding presently are in dispute. Few claim that achievement gains resulting from such expenditures are dramatic. Some might maintain that the schools should not be held accountable for the influx of these "unwanted" funds. If one assumes that all such funds are restricted and that these restrictions preclude the productive use of the funds, then perhaps they should be removed from the cost picture. This charitable course followed, the constant-dollar per capita expenditures and interest for K-12 education decline to $225 in 1950, $315 in 1960, and $520 in 1970.

A second factor influencing increasing costs is a shift in the elementary-secondary school enrollment mix. The Grades 9-12 enrollment was 23% of the K-12 enrollment in 1950 and 1960 and jumped to 28% in 1970. Education at the secondary school level is over half again as expensive as at the elementary school level. Hence, some 3% of the last paragraph's corrected constant-dollar per capita cost for 1970 might be due to that year's proportionately larger secondary school enrollment. If costs are revised accordingly, the constant-dollar per capita costs become $225 for 1950, $315 for 1960, and $505 for 1970. It should by now be evident that the present objective is to say what 1950 education would have cost in 1960 and 1970.

Thomas (1971, Chapter 3) suggests that one possible source of increasing per capita school costs is a change in the output mix. Thus, one might posit that most K-12 offerings in 1950 were of a standard academic type entailing the use of a classroom featuring the usual equipment and printed material. One might posit that, since then, more costly instruction in vocational education and laboratory sciences increasingly has been offered at the junior and senior high schools and that this trend has occurred to a lesser extent at the elementary school level, where such added-cost offerings as orchestra and band are increasingly featured. Had such a shift actually occurred, some of its costs would have been recovered through economies of scale resulting from a combination of factors yielding larger schools and districts in the years since 1950.

For some types of higher-cost output, particularly at the secondary school level, the costs of increasing enrollment can be rather high. Thomas cites a study comparing a public and a vocational high school in the same town during the 1963-64 school year. The per capita cost of vocational education in this instance was 40% higher. Unfortunately for the hypothesis of a cost-unfavorable shift in the output mix, the sketchy extant aggregate data do not conform to expectation--cf, NCES (1975, Table 42). The table shows enrollment in various subject areas by public school students in Grades 7-12 during the years 1948-49, 1960-61, and 1972-73. Natural science enrollments as percentages of total enrollments have increased somewhat--from 58.4% in 1948-49 to 67.0% in 1972-73. Industrial arts enrollments increased from 25.5% to 30.4%. Enrollments in English, Mathematics, Foreign Languages, and Art also increased. Enrollments in Music, Business Education, Agriculture, and Vocational Trade and Industrial Education declined. As the report notes, "enrollment data by subject area show remarkably little variation in the distribution of courses taken over the past 20 years (p. 55)." However, this conclusion is qualified in the next sentence, "Possible changes in the variety and richness of subject offerings are, of course, not revealed by such data."

Federal support for low-achieving components of the enrollment removed, it is likely that no more than 5% of the 1960 expenditures and interest resulted from a cost-unfavorable shift in the output mix, combined with higher-horsepower economics, and that no more than 10% of the 1970 costs resulted from such shifts. That is, 1960 and 1970 figures might be comparable with the 1950 figure for output mix and equipment cost-effectiveness if reduced 5% and 10%, respectively. This done, the resulting constant-dollar per capita costs for K-12 education are $225 in 1950, $300 in 1960, and $470 in 1970.

The figures derived in the last paragraph reference 1950 education for the three factors thus far discussed. Were subsequent K-12 education so structured, these figures assert that maintaining 1950 productivity would require that 1960 per capita achievement units increase by one-third relative to 1950 and that 1970 units more than double relative to 1950.

No basis exists for determining per capita achievement units of any type for any year of the period. The available shreds of evidence--e.g., the ETS study cited in DHEW (1969)--suggest that per capita achievement-unit increases, predicated on 1950 mixes, for 1970 over 1950 probably do not exceed 15%. The scaled-down per capita figure of $470 for 1970 represents 210% of the comparable figure for 1950. One interpretation that can be placed on the 110% cost gain when all other factors are removed is that 15% of this gain purchases gains in achievement, while 95% of the gain rectifies historical exploitation of school personnel.

It is much easier to defend the proposition that school personnel were overworked and underpaid in 1950 than that they were in 1970, are in 1975, or can expect to be in 1980. That is not to say that this component of the workforce has reached a limit for justifying catch-up arguments. The present and foreseeable economic state of the citizenry considered, it simply is going to be more difficult sledding to advance this argument in the future.

The point of this section is neither that past gains of educators in the redistribution of personal income are unjustified nor that additional gains cannot be legitimized. Rather, it is that such gains hereafter in part probably will need be tied to gains in output. Real increases in salaries removed from aggregate direct costs for purposes of studying achievement outputs as a function of direct costs, future real salary gains should be easier to obtain if productivity based on the reduced real costs rises over years.

The costing machinery is going to have to be beefed up to yield the sort of cost allocations guesstimated or invented above and also to disaggregate direct costs by program or achievement domain--e.g., reading as decoding print to speech. Education researchers can contribute to disaggregation of costs by specifying those achievement domains which cost-accounting should reference. However, education economists will have to carry the ball.

Left on their own, economists will expediently favor deferred economic measures--secondary benefits--as outputs on which to estimate educational productivity. Noted above, such analyses--however worthwhile--refer to the productivity of yesterday's educational enterprise, rather than to today's. More current indications of educational productivity are required. The economist who is willing to accommodate more current estimations then faces the dilemma that the universe of discourse for apt achievement measures is in disarray. A goal-directed closing of ranks in this universe is surely required and overdue. It would be an exquisite pleasure to discover that some such ongoing effort as the National Assessment project is responsive to the modelling of scholastic output. The next section discusses NAEP from a standpoint of its possible contribution.

NATIONAL ASSESSMENT PROJECT

It has long been evident that NRTs cannot much contribute to the task of taking stock of national achievement resources. The proprietary firms that develop and market NRTs require a mass market. Until shown an alternative mass market, they cannot be expected to evince much interest in national stocktaking. For these firms, a national stocktaking program would have to be a secondary outcome of income from mass-marketed programs.

DHEW (1969) notes its encouragement of an alternative program for assessing scholastic achievement--the National Assessment of Educational Progress (NAEP), and NCES-funded project of the Education Commission of the States.

> This assessment would involve administering tests measuring standard academic skills to a representative sample of Americans of various ages. Such an assessment, if repeated periodically, would yield for the first time a series of estimates of the change taking place in the intellectual skills and knowledge of the population (p. 66).

Half a dozen years later, NAEP has become a fledgling achievement testing institution and has collected preliminary data in several subject areas--e.g., science, reading, social studies, citizenship. One indication that objectives for a national stocktaking program are not yet in clear view is evident in federal underreporting of data thus far collected by the project--cf, NCES (1975), particularly Table 16, which reduces data on addition, subtraction, multiplication, and division skills to data on computational skills. The project itself is a promising beginning but has several features requiring correction or inviting reconsideration.

- Scope. The scope of NAEP is unduly restricted--to traditional scholastic achievements amenable to paper-pencil assessment--probably because each of the interest groups consulted--scholars, teachers, "thoughtful lay people"--was allowed to veto coverage (cf, Committee on Assessing the Progress of Education, 1969, Chapter 1; NAEP, 1970, Chapter 2). Potential extensions of the common instruction--already featured in the offerings of many schools--into consumership and other coping domains are not included in the project's assessed areas. The project's "thoughtful lay people" for the most part were educated community leaders for whom the skills and knowledges underlying effective adult functioning are demonstrated so reflexively that they pass unnoticed. In light of the USOE-funded Northcutt (1975, in press) study findings, DHEW might eventually find cause to increase the data-collection effort.

- Means of assessment. The widely-used NRTs are practically constrained to favor paper-pencil means of assessment. NAEP data collection is oriented to national stocktaking and so can employ sampling technology--which it does. Thus, the added testing time that might be associated with more frequent departures from paper-pencil testing can be tolerated because spread over more students and schools and because less total data are collected. It is not evident that NAEP takes full advantage of this freedom of action. Written tests generally shy away from the assessment of those fundamental reading skills entailing the decoding of print to speech. They are long on vocabulary and comprehension and short on word-sounding and sentence-reading as decoding-intonation. NAEP apparently is in this tradition.

• Critical prerequisites. NAEP findings for certain proficiencies --e.g., referencing meanings of simpler words--possibly are clouded because a prerequisite proficiency--e.g., word-sounding--is not assessed. One can get around this dependence of one skill upon another--in the illustrative case, by sounding the word to be defined or otherwise characterized. This of course entails departure from the paper-pencil format.

• Domain and score. NAEP assesses skill in addition. Presumably, its addition domain is populated using items or exercises that vary for number of addends, number of digits per addend, carrying requirements, etc., with the range of featured exercises extending from the simplest ones consonant with introductory instruction to the most complex an adult might encounter in the normal course of events--e.g., adding deductions on Form 1040. NAEP tests 9, 13, and 17 year olds and young adults for skill in addition. At each age level, a median percentage correct is computed for responses to the different exercises used to assess addition skill. The finding might be that a median 9 year old provides 85% correct responses; let us assume that this median score applies to the domain's full range, although this might not be the case. If one ignores domain constrictions resulting from discrimination analysis, then an NRT raw score array can be used to obtain the same sort of output description.

Such skills as addition are defined on increasing problem complexity. Addition problems increase in complexity along two or more dimensions. If one must strike a balance between reporting nit and fuzzy breadth, then addition problems can be unidimensionally scaled for complexity on the basis of outcomes. Given such a scale, one begins by sampling problems at not-too-widely separated points along it. Imagine that Point 0 references no problems and represents an utter lack of proficiency for addition; Point 1 references problems requiring addition of two one-digit addends without carrying; Point 2 references problems requiring addition of three two-digit addends without carrying; Point 3 references problems requiring addition of three multidigit addends with carrying; Point 4 is a next higher point; Point 5 is the top of the scale.

If a test referencing such an analytic domain then is administered, it becomes possible to say, for example, that 1% of 9 year olds are at Point 0, 9% at or just above Point 1, 25% at or just above Point 2, 50% at or just above Point 3, 14% at or just above Point 4, and 1% at Point 5. Findings for the 9 year old cohort then can be reported as a "frequency x complexity" function. Where NAEP might report a median attainment of 85% referencing an undetermined and, indeed, indeterminate point on the complexity scale, the entertained alternative is to report the values 1%, 9%, 25%, 50%, 14%, 1% for an age-level function of problem complexity and to compare functions at different age-levels and across years for the same age-level in all of the usual ways. Although it costs more to place given fiducial limits around a

function than around a point and uses a bit more space to report findings, the additional informative power appears to warrant these minor increases in cost. The mandating of minimal outputs cannot be enforced without such information, which might be required to show 90% of 13 year olds at Point 4 for addition and 10% at Point 5. Meanwhile, NAEP domains are almost as fuzzy as those employed by NRTs.

• Geographic reference. Consonant with guidance afforded by school officials at the time the project got underway, NAEP reports no basis for making achievement comparisons between and within states. The same officials now criticize the project--appropriately of course--for this "oversight" (Maeroff, 1975). The moral has perhaps been too pointedly drawn by Wynne (1970), who comments on a possible constituency for education researchers. If the states wish to zero-sum game in an absolute achievement context, there can be no harm in this. The outputs of common instruction will not overnight reflect attainment of useful minimal standards for achievement. Geographic differentials on progress toward such an objective might be of interest. Outside the areas of common instruction, some achievements might prove of interest to a national stocktaking program. In these "elective domains," geographic differentials might occur forever. All things considered, national stocktaking for achievements should reflect geographic subdivisions.

• Primary-secondary effects identities. Entertainably, some primary effects of common instruction simply are bridges to securing other primary effects. Most primary effects, where retained, show up in adulthood as simple secondary effects in a different or wider context. Worth considering is the proposition that the consumer and other adult coping proficiencies of the genre studied by Northcutt (1975, in press) should be viewed as primary effects when referencing children in school and secondary effects when referencing young adults. The schools are moving toward extending the common instruction to pick up many such proficiencies. Were the scope of national stocktaking broadened to accommodate adult coping proficiencies, one could gauge how quickly the common instruction is being extended to encompass these proficiencies. Aside from a rather abstract handling of citizenship proficiencies, NAEP follows the maxim of the organized interests that, judged on the character of life outside the schools, schooling is wonderland. Perhaps it should be less harsh and more protected. But it also should respond to findings such as Northcutt's with some of the sense of concern regarding the findings expressed by Education Commissioner Terrel Bell (Education Daily, October 30, 1975).

Many of the deficiencies of NAEP are those of NRT efforts--not too surprising since NAEP relied heavily on NRT firms both during project formulation and subsequent data collection. All such deficiencies are rectifiable--either under NAEP's tent or some other. This paper does not attempt to develop a national achievement testing program per se. Instead, it specifies those features of such a program which serve stocktaking objectives while tying the program

to scholastic information needs of parents and teachers. Comments are meant to apply only to primary effects of common instruction because specification and assessment of the primary effects of other instruction might pose special problems. The common instruction for the most part teaches basic proficiencies in such traditional content areas as reading, mathematics, and language skills and in such other areas of recent interest as consumership, citizenship as rights complementing obligations, and operating effectively in economic, social, and aesthetic space.

One should not lose sight of the fact that some carriage manufacturers made the transition to manufacturing automobiles--not because they loved carriages less but because the mass market came to love their carriages horseless. It is perhaps a common dilemma for firms marketing NRTs and those who would render achievement testing more than ego-gratifying to rich school districts that the mass market currently buys test batteries primarily to engage once annually in the national educational zero-sum game. Rather effortless gains in educational productivity probably can be expected when those who develop mastery achievement tests and readily-used, cost-effective means for automating the frequent use of such tests to assist instructional management deliver on a decade of promises. In so doing, they will illuminate a less frivolous mass market and so might persuade some espousers of pterodactyloid principles to make the necessary adjustments.

THE MASTERY CONCEPT

Consonant with Kuhn's (1962) law for the conditions under which a prevailing paradigm will give way, NRTs will not fall into disuse simply because attention is focused on their warts. Nor will NRT advocates be won over by conciliatory postures--cf, Harris et al. (1974), an "under one tent" account which casts mastery achievement test advocates as the dedicated males of black widow spider matings with NRT advocates. The prevailing NRTs will be superseded when alternative tests having their few advantages and avoiding their many defects are available and command respect in a mass market.

An MAT perspective is required that ties national stocktaking to instructional proficiency domains and classroom instructional management to similar but typically narrower domains, while leaving the details of offered common instruction under local control. Justified on purely pedagogical grounds, such a perspective provides resourceful NRT publishers with a replacement mass market--that for instructional management--and a chance to express their gratitude by rendering the public service of making minor resource contributions to a national stocktaking program.

The development of tests enabling responsive instruction and achievement stocktaking has been underway for a decade or more.

Such tests can be distinguished on several grounds--e.g., domain- versus criterion-referenced (DRT versus CRT), specific to a particular instructional program versus specific to a particular skills domain. The common feature of these tests is an orientation to mastery of specified knowledges and skills. Whatever their differences, such tests class under the general heading of "mastery" tests.

Those antagonistic to the mastery concept sometimes impute to its users a naive view of reality which occasional advocates might indeed possess. However, the notion of absolute achievement grounding mastery assessment is fully operationalizable or nearly so and is perhaps the only useful achievement concept at this stage. Relativity in physics became tenable only because an absolute framework existed within which the empirical meaning of relativisms could be established. Relativity in achievement is a rubbery concept which either does not aspire to have empirical meaning or is predicated on some new logic not yet read into the public domain. Herein, the mastery concept simply signifies a decision process that renders it possible to say a given student probably has or has not attained a proficiency whose acquisition given instruction intends, with the consequences less than catastrophic if such a decision is occasionally wrong.

Another argument of orthodox thinkers is that the mastery concept more often than not does not apply to the more significant instructional intents of elementary school offerings. Two responses to this argument are pertinent. First, even though skills analysts find the going easier in some instructional domains than in others, the alternative quasi-G-factoring of proficiency simply is unutilitarian. Skull contents resulting from vacuum-cleaning the environment might serve television game show requirements. Their explication, if it can be called that, does not much lighten the teaching load. Second, the argument is an empirical assertion whose merit eventually will be decided on the basis of extent to which mastery-testing advocates prove the concept's utility.

It might often be true but is not particularly useful that NRT domains are too broad, CRT domains to nitty, and DRT domains just right. The achievement testing moral is more complex than that of a Goldilocks story. The remaining sections provide a conceptual basis for distinguishing useful forms of DRTs and CRTs.

CRTs AND DRTs AS CLASSES OF MATs

The earlier mastery achievement tests all were denoted criterion-referenced tests (CRTs). The items of a CRT reference a learning or proficiency domain--or, more precisely, a stimulus domain on which a particular proficiency has been defined. An illustrative stimulus domain is high-frequency CVC (consonant-vowel-consonant) words--e.g., cap, leg, miss, shot, bun, . . . Alternative proficiencies can be defined on such domains--e.g., word-sounding, word-spelling, word-writing. Consider a CRT for word-sounding defined on a CVC domain.

An item which asks a student to sound cap typically implies an absolute proficiency standard or criterion, which an offered response either does or does not meet. Where the domain is homogeneous in the sense that all items are viewed as exemplarizing a specified proficiency and a CRT features several such items, the criterion used by a teacher to decide whether to advance or shift a student to other instruction--e.g., word-sounding defined on a CVCe domain --usually entails less than errorless performance on the test. Where the proficiency criterion is absolute, one might say that this criterion is slightly discounted to a mastery or performance criterion which takes into account the eventuality that noise factors occasionally mask proficiency. Thus, a CRT administered at some point in word-sounding instruction might ask a student to sound CVC words in written form and entail advancing the student to a next portion of the instructional sequence only if at least 90% of his word-sounding responses are correct. If such testing is sufficiently frequent and tied to ongoing instruction, occasional wrong decisions concerning a student's status for a specified proficiency are acceptable.

To promote the explicit specification of domains and systematic sampling from domains during test formation, such writers as Hively et al. (1968) and Hively et al. (1973) introduced the notions of item universes, item domains, and stratified item domains. While item terminology was used, the referenced universes and domains also are stimulus domains on which specified proficiencies are defined. At about the same time, multiple matrix sampling of domains entered the conceptual space for mastery achievement testing (cf, Shoemaker, 1973). Although the primary aim of such efforts was to improve the technical basis for criterion-referenced testing--primarily with regard to domain and test formation--the broadening of domains relative to those for CRTs is inherent in these efforts and their exploitation in the schools. Since 1973, the notion has emerged of domain referenced tests (DRTs) as more-general alternatives to CRTs as mastery achievement-testing devices.

Thus, for example, the National Countil of Teachers of English (1975) characterizes NRTs as broad or fuzzy for domain, states that "criterion-referenced tests divide the world of English into tiny fragments of learning" (p. 17), and goes on to note that

> Domain-referenced tests were created to strike a balance between fragmentation and fuzziness. A test-maker defines a domain of learning and criteria for success within that domain. A domain-referenced test in literature, for example, could deal with the ability to recognize and discriminate among the common types of figurative language rather than to recognize metaphors (pp. 17-18).

Prior to 1973 or 1974, it could be said that DRTs are a technically-improved class of CRTs. As NCTE (1975) indicates, the emerging convention is to view DRTs and CRTs as scope-distinguished classes of MATs. This paper envisions DRTs and CRTs which are equally acceptable on technical grounds. It subscribes to the emerging convention, but finds the convention insufficient for purposes of distinguishing DRTs and CRTs.[2]

INSTRUCTIONAL HIERARCHIES

Two sorts of charting structures are useful to characterizing instructional programs:

- A taxonomic hierarchy, typically featuring single classification, illustrated by "Catalog of Objectives, Sobar Reading" (Science Research Associates, Inc., 1974).

- A skills hierarchy, favored by skills analysts--e.g., Gagné (1970)--and many instructional developers.

Were it necessary to characterize and distinguish NRTs, DRTs, and CRTs just in terms of the levels of a taxonomic hierarchy, the differentiation would be as illustrated in Table 1. Unfortunately, the guidance afforded by such a characterization tends to be both fuzzy and misleading--the basis for the earlier comment that scope-distinguishing of DRTs and CRTs is insufficient. A clearer picture of differences in form and applicability of DRTs and CRTs is afforded by considering the two types of tests in the context of both types of hierarchies.

Both hierarchies assume that entering students have certain entry skills. Such skills are posited to render the scope of instruction finite. Apt to both hierarchies is an apical term whose significance is overall breadth of the intended skills domain. The apex term of a taxonomic hierarchy encompasses one or more culminant skills of intended instruction; these skills go unnamed in the taxonomic hierarchy. Conversely, the culminant skills encompassed by a skills hierarchy are specified in the hierarchy. A taxonomic hierarchy is useful for inventorying post-entry elemental skills; a skills hierarchy, for identifying assembly skills predicated on elemental skills and for forming an instructional sequence of elemental and assembly skills.

---

[2]Nomenclaturists have entertained themselves and others by placing various words in the blank of "____-referenced tests." Present remarks find DRTs and CRTs as herein defined sufficiently exhaustive of the MAT universe.

Table 1

Distinguished and Assessed Taxonomic "Levels of Instruction"

Taxonomic Level	Level Descriptor	Characteristic Test at Level	Exemplars	
1	SUBJECT AREA	None	READING SKILLS	
2	SUBJECT SUB-AREA	Norm-Referenced	WORD ATTACK SKILLS	VOCABULARY SKILLS
3	SKILLS DOMAIN	Domain-Referenced	CONSONANT SOUNDING SKILLS	WORD MEANING SKILLS
4	SKILL	Domain-Referenced or None	CONSONANT BLENDING SKILLS	CONCRETE-NOUN DEFINING SKILLS
5	OBJECTIVE	Criterion-Referenced	BEGINNING CONSONANT BLENDS: An oral word featuring such a blend presented, from a set of alternatives select a written word featuring the same blend.	ILLUSTRATED INTENT FOR CONCRETE NOUNS: Such a noun presented in written or oral form, from a set of alternatives select an illustration showing its intent.
6	INSTRUCTIONAL ELEMENT	Criterion-Referenced or None	BEGINNING XL AND XR TWO-LETTER COMBINATIONS (SOME 27)	THE MORE-COMMON ANIMATE COUNT NOUN COMBINATIONS (SOME FEW K)

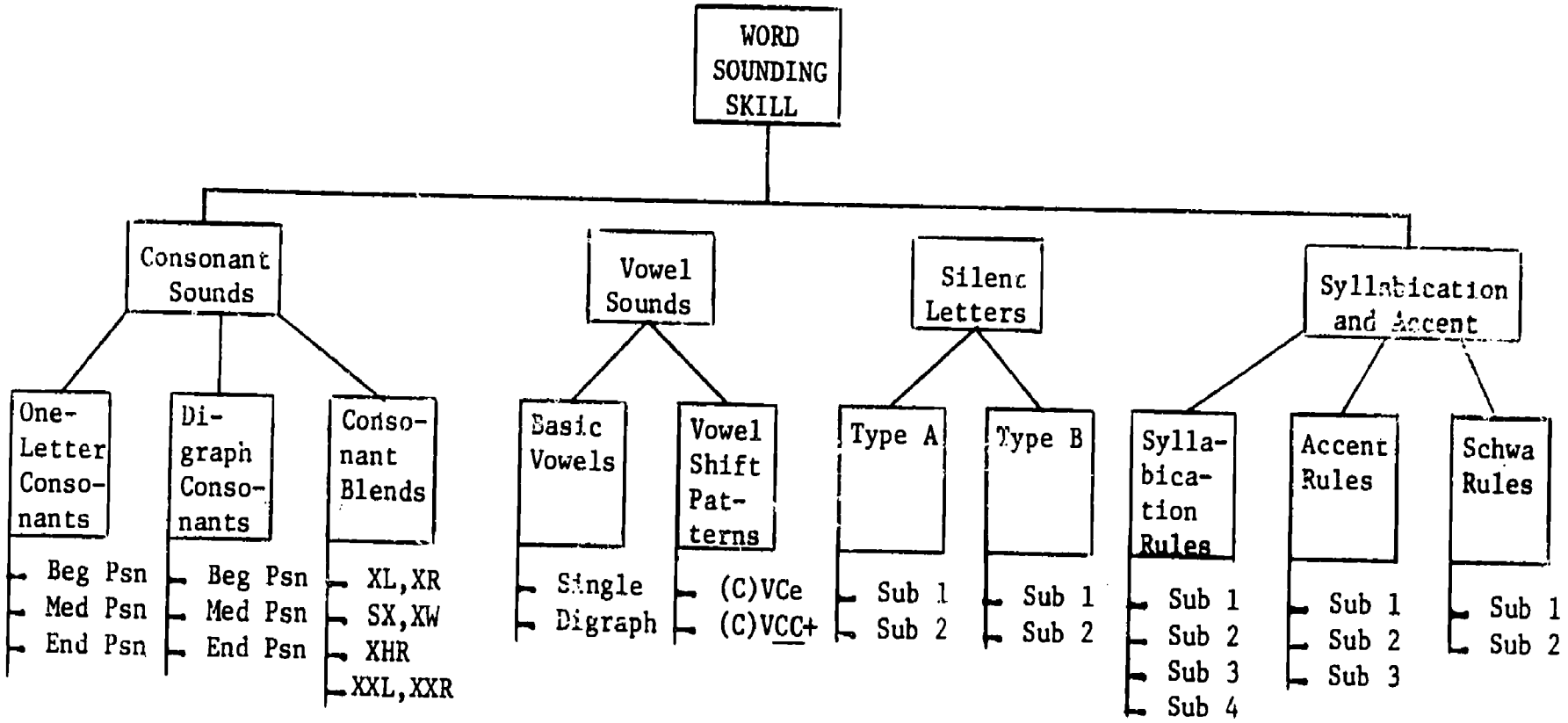

Figure 1. Illustrative Taxonomic Hierarchy. Actual instruction implicated by any such structure typically references terminal nodes of the hierarchy. Terminal boxes--e.g., Single Consonant Sounds in Beginning Position--reference elemental skills. This diagrammatic form is not useful for showing the assembly skills which result through combination of elemental or lower-level assembly skills.

Taxonomic Hierarchy

The terminal nodes--collectively, the bottom line--of an inverted-tree diagrammatic form for the taxonomic hierarchy inventory those elemental skills to be introduced by intended instruction. Rubrics intervening between the apex and bottom line typically serve only as intermediate subsumers of bottom-line elemental skills. This is not so where concept induction is at issue. Outside of instruction in concept learning, the concept-induction implications of higher-level rubrics of taxonomic hierarchies typically are of little instructional interest. A taxonomic hierarchy for word-sounding (or word-attack) skills is illustrated in Figure 1. Its higher-level rubrics apparently have no more than incidental concept-induction implications for word-sounding instruction.

The bottom line of a taxonomic hierarchy inventories elemental skills to be introduced by intended instruction. The order in which these instructional elements are introduced typically is not specified but is specifiable--e.g., through sequential coding of terminal nodes. More importantly, the hierarchy fails to specify the assembly skills intervening between elemental and culminant skills of intended instruction. A taxonomic hierarchy is an insufficient basis for describing a progression of instructed skills.

Taxonomic hierarchies are means for inventorying elemental skills of intended instruction, particularly if breadth of terminal nodes--which need not all be at the same level--is conditioned by a pragmatic criterion for extent of instruction. An illustrative such criterion might require an elemental skill at a specified instructional level to entail on the average at least one but not more than three hours of instruction. Such a criterion fends off analytic trivialization of elemental skills--the problem posed by a potential infinite regress for analytic depth.

Skills Hierarchy

Entry skills are culminants of previous instruction or experience. An apex-level skills domain specified and elemental skills derived using a taxonomic hierarchical structure, an instructional design and development effort then must:

- Preliminarily sequence the elemental skills.
- Identify assembly skills which combine elemental skills, lower-level assembly skills, and combinations of elemental and lower-level assembly skills.
- Settle on an overall sequencing of skills.

Any such effort yields a skills hierarchy. Such a hierarchy reflects the relative interval in time at which each elemental skill is instructed, the assembly skills deriving from instruction featuring previously acquired lower-level skills, the antecedents to each assembly skill, and the relative time at which each assembly skill is instructed. The hierarchical status of an assembly skill can be distinguished on the basis of the hierarchical status of its antecedents. Hence, no information is lost if the diagrammatic form of a skills hierarchy is that of a two-level flowchart whose x-axis is a relative time line. A simpler such structure is shown in Figure 2. In such a structure, elemental skills occur at the bottom level; assembly skills, at the top level. Figure 2 is illustrative for content and does not assert that phonics-based word-attack instruction should be initiated as illustrated.

Unlike a taxonomic hierarchy, each box or node of the elemental and assembly skills levels of a skills hierarchy references actual instruction of a unique skill. Also unlike a taxonomic hierarchy, a skills hierarchy is exhaustive in the sense that no instructional gaps occur between any two points on the relative time line.

The apex-level subsumer shown in Figure 2 is a feature of a taxonomic hierarchy. Actual instructional programs--e.g., for earlier reading--are more complex than that for the word-attack program illustrated in Figure 2. Consider sentence-reading skill as subsuming word-attack and sentence-intonation skills. An extended skills hierarchy handles this situation for purposes of sequencing instruction.

Extended Skills Hierarchy

Word-attack instruction is not treated as an exhaustive block to be completed before other instruction underlying reading skill is initiated. A child typically is asked to learn to read simpler sentences with appropriate intonation and to characterize intent of words and sentence forms encountered shortly after earliest word-attack instruction. When skills falling under vocabulary, comprehension, and other rubrics are intermingled on the instructional time line with those falling under the word-attack rubric, two-level flowcharting becomes more cluttered with arrows linking antecedents to assembly skills. Such interlarding of skills domains does not invalidate the propositions that each box at the elemental and assembly skills levels references actual instruction of a unique skill and that explicated instruction between points on the instructional time line is exhaustive. Indeed, it introduces interskill assembly instruction where needed.

Consider a skills hierarchy for sentence-reading as extended. Imagine that a small amount of word-attack instruction--e.g., addressing VC and CVC words--and a small amount of sentence-intonation instruction addressing simpler sentences are interlarded. For illustrative purposes, assume that these interlarded instructional sequences

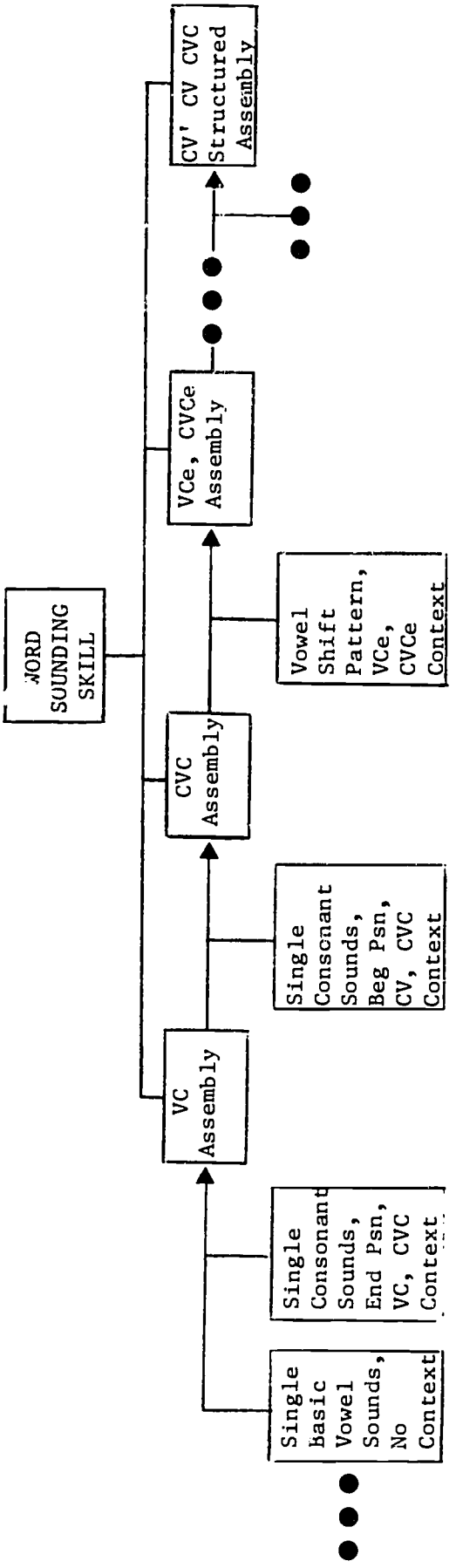

Figure 2. Illustrative Skills Hierarchy. Each box at the elemental and assembly skills level represents actual instruction. The flowcharting scheme adopted obscures the fact that assembly skills are at progressively higher levels for the particular portions of the particular instruction illustrated. Whether given assembly instruction features as antecedents instructional effects referencing two or more elemental skills boxes, one elemental skills box and one assembly skills box, or two or more assembly skills boxes, each assembly skill is an instance of word sounding skill.

culminate on word-attack and sentence-intonation skills which are transformed into sentence-reading skill only in consequence of inter-skill assembly instruction. Such instruction addresses a higher-order assembly skill and culminates on sentence-reading skill referencing simpler sentences using VC and CVC words. Actual instruction over extended time yields assembly skills of higher and higher order. Some of these higher-order assembly skills probably need not be formally instructed; given the parts, the student will put them together on his own. Some do require formal instruction. An extended skills hierarchy marks these assembly skills. A testing program might recast the extended skills hierarchy to better reflect progressive assembly. A synthetic hierarchy might be used for this purpose.

Synthetic Hierarchy

The illustrative extended skills hierarchy features a sentence-reading skill that is a higher-order assembly whose antecedents are word-attack and sentence-intonation skills over restricted ranges. Actual instruction over extended time yields many such first-higher-order assemblies, some second-higher-order assemblies whose antecedents are first-higher-order assemblies, some third-higher-order assemblies whose antecedents are second-higher-order assemblies, etc. Not each "logically-derived" such higher-order assembly skill requires instruction. Those that do not do not appear in an extended skills hierarchy. Those which do appear invite testing. One can distinguish levels (and breadths) of testing by representing instruction using a synthetic hierarchy which combines certain features of the taxonomic and skills hierarchies.

Again using the illustrative extended skills hierarchies, the synthetic hierarchy places the skills hierarchies for restricted word-attack and restricted sentence-intonation instruction side-by-side at the same level. These hierarchies are subsumed using an assembly skill for sentence reading. This is the simplest synthetic hierarchy. It can be extended to the right and upward to whatever extent circumstances warrant. In it, one gives up the relative time line of the extended skills hierarchy, while overcoming the silence of taxonomic hierarchies concerning the assembly components of instruction. The synthetic hierarchy simply is a transformation of the extended skills hierarchy, a useful form of representation when one wishes to portray the increasing breadth of tested domains.

Where a network of developing proficiencies can be expressed using an instructional hierarchy, the nittier--not a pejorative term here--CRT programs reach down to the elemental skills level of a skills hierarchy. The less-nitty CRT programs reach down to the lowest assembly skills level. Programs of both types exhaustively test for instructional effects for every skill portrayed at and above the lowest level selected for testing. Conversely, a DRT program used for state and national stocktaking purposes seeks to place the

respondent along a stimulus-referenced complexity scale--e.g., VC, CVC, CVCe, . . .--for specified proficiency--or to be able to say that a given student has mastered the sounding of CVC words; a second student, the sounding of all two-syllable words; etc. Such scales are reflected as assembly skills progressions--cf, Figure 2.

KNOWLEDGE DOMAINS

Taxonomic knowledge hierarchies are common and have conceptual meaning. However, knowledges featured in the common instruction of elementary schools tend to occur in single cells of such hierarchies. It might be possible to define knowledge in a specified subject domain on increasing extent of the domain. Thus, one can increase the scope of political geography by adding surface per se, by adding superordinate categories such as continents and hemispheres, and by adding subordinate categories such as counties. However, there is no accepted way to expand a particular knowledge domain. It is neither pedagogically compelling nor good politics to use an arbitrary model for expanding a given knowledge domain just to serve state and national stocktaking objectives.

One alternative is to distinguish a small number of pertinent scales for knowledge and to obtain percentage-score distributions for each age cohort referencing each such scale. It might be convenient to block percentage scores by range--e.g., using some six ranges--for reporting purposes. For each scale for each cohort, then, data might be reported as a "frequency x score range" distribution. More analytic solutions should be pursued.

Two examples of the envisioned approach are presented below. Consider first the political geography of the United States. For present purposes, imagine that three aspects of instruction are distinguished:

- Verbal knowledge. This might entail classifying 206 names --2 bordering countries, 4 bordering bodies of water, 50 states, 50 capitals, and 100 principal cities--by category.

- Relational knowledge. This might entail placing on an outline map--showing the pertinent state and national boundaries, with stars representing state capitals and dots representing other principal cities--labels for the states, capitals, other principal cities, and bounding countries and bodies of water.

- Quantitative knowledge. This might entail placing on a fully-labelled outline map the area and population values for states and population values for cities.

The portrayed and antedating verbal knowledge might be of concern to a CRT program assisting instructional management, but probably is beneath state and national interest for policymaking purposes. It

is possible to teach either the relational or the quantitative knowledge after the verbal knowledge is acquired. Hence independent scales for relational and quantitative knowledge might be distinguished. Each of these scales references a domain of U. S. political geography. Since neither domain--as illustratively defined--contains all that much information, testing might be exhaustive, with some students at each age level in each distinguished geographic area responding to "items" referencing a portion of the domain and data reported as "frequency x score range" distributions.

The ranges used should be a function of chance probability for a correct response. Where chance probability is rather high--e.g., .20 for five-choice, selected-response items--Level 1 (0-20% raw scores) might be interpreted as indicating a lack of proficiency and Level 6 (95-100%) as indicating mastery. Where chance probability is low--the case for a relational knowledge scale featuring the placement of 206 labels on a map--Level 1 might be restricted (e.g., to the range 0-5% raw scores) and Level 6 expanded (e.g., to the range 90-100%). The interpretation of data distributions will be made easier if only a few--e.g., Type 1 and Type 2--range sequences are used. Illustrative data for the relational knowledge of U. S. political geography of 9-year-olds then references the Type 2 six-level range sequence and might take the form: 1%, 4%, 10%, 50%, 30%, 5%.

In passing, the prevailing NRT programs must make a big thing of test security because their tests feature handfuls of items which, at great expense, are found to get to the heart of the factor-analytic view of cognitive proficiencies. Test security for the envisioned DRT program simply comes down to not announcing to particular classroom teachers or their students just what tests a selected class will negotiate for stocktaking purposes until a day or so before the tests are administered. The explicated domains for all tests of the DRT battery should be in the public domain, and publications fully describing these domains should be in every classroom. So long as a class to be tested for relational knowledge on political geography is not practiced exclusively in that domain just prior to testing, there can be no harm in fully describing such a test along with a hundred companion tests, any one of which might be administered to the class. Not only should the fully described pertinent domains and testing procedure referencing these domains be readily available to the public; the public should be encouraged to examine these materials and offer criticisms where some elements of coverage seem pointless or overly emphasized, some potential elements appear uncovered or undercovered, etc. Education will be well-served by stripping away the mystique which now surrounds testing. The envisioned DRT program should permit whatever level of scrutiny any individual might wish to give it.

Citizenship is a second knowledge domain in which productivity of the schools might be judged. It is apparent that the common instruction in citizenship cannot have the objective of producing experts in the pertinent constitutional and statutory law. Rather, it seeks to inform the individual concerning his basic rights and obligations.

The reasoning underlying obligations and--to a lesser extent rights--of citizenship is extended. The obligations and rights themselves are few in number.

Just as one can imagine a knowledge test for world political geography "riding above" that sketched earlier for political geography of the U. S., one can imagine two levels of testing for citizenship. The lower-level test might reference listings of obligations and rights and a few widely-understandable antecedents and consequences of the cited obligations and rights. If one uses that elastic measure--the "fact"--there probably are no more than five hundred citizenship facts one should know to function effectively for citizenship during adulthood. Some half of these might be featured in a lower-level test domain, with the domain perhaps further differentiated into obligation and rights domains and with knowledge in each domain tested by a DRT. If the tests which result when domains are exhausted--e.g., 100 items--are too long to be negotiated in their entirety, then sampling designs permit exhausting the domain while administering fewer items to given students. For each scale--e.g., obligations and rights at Levels 1 and 2--data by age cohort might be reported as a "frequency x score range" distribution.

CRTs referencing knowledge domains assess proficiency for progressively-introduced portions of the knowledge domain. An indication of the alternative means for differentiating CRT domains probably can be obtained by examining a sample of pertinent textbooks. It might often prove the case that CRT domains referencing a specified content or knowledge domain tend to be similar for coverage from one textbook to the next, with only order of coverage varying. Whether CRTs referencing knowledge domains must be textbook-specific or can be more generally specified is left open here.

CRT AND DRT ROLES

This paper assigns to CRTs the role of assisting management of day-to-day instruction. A CRT program might address each skill distinguished in an extended skills hierarchy. Conversely, the program might address only the assembly skills of simple skills hierarchies and the higher-order skills of synthetic hierarchies. The level of criterion-referenced testing is negotiable; whether assessed skills are nitty or less nitty is a local option. This also is true for knowledge domains, although the pertinent "hierarchies" in most instances probably will prove to be single-level.

Consider a CRT program's lowest tested level in a synthetic hierarchy the program's base level. It is a local option concerning the base level at and above which instructional management information is useful. One school might find it useful to define the base level on elemental skills; another, to define the base level on assembly

skills of simple skills hierarchies. Whatever the base level, all skills at and above that level in a synthetic hierarchy probably should be assessed as pertinent instruction is completed.

This paper assigns to DRTs the role of periodic assessment of knowledges and level-conventionalized skills for state and national achievement-stocktaking purposes. The domains of envisioned DRTs should apply to stocktaking objectives, whatever the details of local common instruction. One such DRT--for skill in addition--was illustrated in an earlier section. The illustrative DRT was formulated in terms of an outcome-defined six-level unidimensional scale for problem complexity. In it, proficiency at each level has a fully specifiable characteristic meaning. Lowest-level proficiency in the illustrative test constitutes no addition proficiency whatsoever. Highest-level proficiency reflects as much skill as any adult might require to successfully accomplish those common undertakings of adulthood entailing addition skill.

A stocktaking program has the options of specifying how students will be matched to levels of a test or allowing a participating school to effect matching. If matching is program-mandated, one approach might be to administer a Level 1 test to an nth of the tested sample (e.g., a sample of fourth graders or 9 year olds, geographically stratified), a Level 2 test to a second nth of the sample, . . ., and a Level N test to a final nth of the sample. Such an approach has the undesirable effect of wasting the time of too many students on problems which are too far above or below their proficiencies. It also strings out the inferential machinery which must be explained to users.

Where a school uses envisioned CRTs or some other means that equip its teachers to make good bets concerning the level in a testing sequence at which a given student is proficient, teachers should be able to so match students to levels of a pertinent DRT as to minimize wasted student time and to render explanations of the inferential machinery employed less cumbersome. If a stocktaking program designates a given class, grade, or age cohort in a given school for participation in a given study of proficiency, informed teachers should be able to predict rather well the level at which each participating student is proficient. Matching procedure then might entail administering tests at the estimated proficiency level and the next-lower and next-higher levels. Such matching might be considered legitimate if outcomes are as follow:

- Errorless performance, suitably discounted for noise effects, occurs at the lowest level tested.

- Less than errorless performance, suitably discounted for noise effects, occurs at the highest level tested.

Where either of these criteria is not met by student responses offered during initial testing, the stocktaking program might either

require additional testing of the student or the testing of an alternate student and data substitution.

The assembly skills level of Figure 2 illustrates a scaled domain for word-sounding skill. DRTs assessing fundamental skills are believed usually to reference such domains, which scale for stimulus or problem complexity. Illustrative scaled domains for word-sounding and vocabulary skills are discussed in the next section.

ILLUSTRATIVE SCALED DOMAINS

The basic skills of common instruction often can be defined on scales of increasing stimulus complexity. Subtraction, multiplication, and division problems increase in complexity much like addition problems considered earlier. Problems of mixed arithmetic increase in complexity along several dimensions but can be unidimensionally scaled for difficulty, like the simpler problems. Measurement problems can be scaled in terms of measurement scale gradations and number of operations inherent in problem solution. Most basic skills of mathematics appear amenable to assessment along scales for increasing complexity, with problem or task structures at each level straightforwardly describable. In consequence, saying that 40% of the 9-year-old cohort is proficient at an ith level of any such scale implies a class of problems or tasks which these children have mastered.

Present remarks assume that proficiency defined on a scale of increasing complexity entails that an individual who is proficient at an ith level is proficient at all lower levels. When occasions arise wherein this is untrue for more than a small proportion of students, testing will need be more extensive and data summarization and reporting will need depart somewhat from the paradigmatic form featured in this paper. The issue is not whether a Guttman-type orientation to scale characteristics is univerally warranted. Where warranted, data collection, summarization, and reporting should reflect such scales. Where not, pragmatic adjustments will need be made.

Word-Sounding

Reading and other communication skills areas of instruction reflect many skills referencing complexity-scaled domains. Two illustrations are word-sounding and vocabulary skills. Whatever the details of instruction for such skills, their skills hierarchies tend to reflect successive assemblies which are of increasing complexity along one or a few dimensions. Illustrative levels for word-sounding are:

- Level 1. VC words (e.g., as, ebb, it, odd, us), CV words (e.g., me, so; la, the, do, by).

- Level 2. CVC words (e.g., bath, shed, mill, log, mush; bar, Cher, nor, bush); CCVC words (e.g., trap, clam, span, staff, shred); CVCC words (e.g., pact, part, pant, past, pelt, thank; calf, kiln).

- Level 3. VCe words (e.g., ate, eve, ice, ode, use; are, ere, ore); CVCe words (e.g., late, mice; share); CVC words (e.g., gain, taut, dead, look, mouth; great, tooth); all remaining one-syllable words.

- Level 4. All two-syllable words (e.g., Ba'bel, ra'pid; ba'sal, ra'ven; canal', repell'; matter, platter).

- Level 5. All three-syllable words of Latin origin (e.g., article, delegate, united, vacillate; ravenous, carbonic, transference, vacation).

- Level 6. All remaining word structures of the lexicon pertinent to effective adult functioning.

If assembly skills for word-sounding are grouped about as they are here, then teachers should be able to assign students to levels, regardless of the details of local instruction for word-sounding. Whether instruction is phonics-based or "sight-syllabary-based," the levelling of an apt DRT should do justice to the local instructional effort. If the illustrative levelling does not conform to this requirement, some slight modification of it should.

A useful stocktaking test for word-sounding skill identifies the highest level in a test sequence referencing a complexity-scaled domain at which an individual is proficient. For each age cohort tested, data are reported as a "frequency x levels" distribution. A spelling test sequence well might use the same lexical domain and the same levels, with the DRTs differing only in presentation-response characteristics. Word-sounding words require written presentation and oral responding; spelling words, oral presentation and written or oral responding (which, of course, define two somewhat different skills).

Vocabulary

Whether the context is an intelligence test, an NRT, or a DRT and whether vocabulary skill is defined in terms of usage or illustration-definition, testing involves using a small sample of words to reach inferences concerning how many words the respondent commands relative to a larger universe. The larger universe typically is viewed as all of the words in an unabridged dictionary. Using a sample of words to estimate proficiency referencing an extended

universe entails modelling the universe along one or more dimensions --e.g., frequency in print, frequency in speech (apt but hard to come by), morphological complexity, semantic complexity.

Present interest centers on a nonrepresentative domain of the universe for English words--a lexicon containing the more pertinent words to effective adult functioning. Imagine that the lexicon consists of 20K words. By comparison, desk dictionaries contain 50-100K entries. Various techniques exist for unidimensionally scaling these words for order in which their usage or meaning comes to be understood on the average. Assume such scaling has occurred. Illustrative levels for vocabulary skill are:

- Levels 1-2. The lowest-scaled two successive 1000s of words.
- Levels 3-4. The next two successive 2000s of words.
- Levels 5-6. The next two successive 3000s of words.
- Levels 7-8. The highest-scaled two successive 4000s of words.

Extent of the lexicon and levelling are illustrative. It is assumed that all adults should be proficient for all words on the list. Entries are differential across illustrated levels to insure finer stocktaking at the lower end of the scale.

The list--as a common dictionary available to all--should be published as scale-levelled "word + characterization" entries. The published list constitutes a public basis for interpreting state and national data in "frequency x levels" form, by age cohort. The common dictionary should be prominently featured in all classrooms and should be available to parents at the lowest price at which it can be offered. The dictionary should, of course, be updated from time to time as requirements for effective functioning in adulthood change.

Entries of the common dictionary suitably scaled, a school employing a CRT program to assist instructional management decisions should be able to place students along the vocabulary scale sufficiently well to insure legitimacy of three-level testing--the level of predicted proficiency and the next-lower and next-higher levels. Different subgroups of an age cohort sample might take different 10-20 item tests at a given level, with the totality of tests at a level either exhausting the domain's words at that level or constituting such massive coverage as to dispel the notion that sampling such "heterogeneous" levels might have different consequences based on different samples. Vocabulary skills incorporate a variety of conceptual skills and are sufficiently important in any scheme for reporting the nation's achievement resources to warrant disproportionate data-collection expenditures if this is required to reach good estimates.

Teaching-to-the-test problems arise only when the test features a recurring small sample from a large domain or universe, so that inferences referencing the domain or universe are invalidated if strict test security is not maintained. As noted earlier, there is no need for such security in envisioned DRT programs, since they exhaustively describe the pertinent domains and make this information available to the public. Imagine that some student on his own or with teacher assistance so crams before taking a few small tests in unknown areas by acquiring many of the proficiencies advertised in public accounts of a hundred potentially pertinent domains that he manages to improve tested performance relative to what it otherwise would have been. The common instruction could only profit from such zeal.

If "reading comprehension" is a useful term, it must not reference proficiencies assessed elsewhere by an exhaustive stocktaking program addressing basic proficiencies. The particular words used in reading comprehension exercises are assessed for comprehension by DRTs for conceptual-vocabulary skills. Particular knowledges are assessed by DRTs referencing pertinent knowledge domains. The term apparently signifies information-processing skill for prose material that varies for syntactic extent and complexity. Given an intersentence grammar that predicts the difficulty for obtaining implications of arguments delivered using constructions of a class, reading comprehension skill might straightforwardly be defined on a scale of increasing complexity. One can proceed more intuitively. Existing tests either reflect such a scale based on intuitive analysis or masquerade as useful tests. If useful, other broad rubrics--e.g., critical reading skills--also should reference one or more scales of increasing complexity. Entertainably, such skills designations can be rendered useful within a DRT framework by stripping away some of their mystery. They probably are higher-level assembly skills whose assessment at their levels need not recapitulate the totality of skills lying at lower levels in the grand synthetic hierarchies they crown.

WHITHER THE SCHOOLS

As with many other concepts he has introduced to enrichen education discourse, one cannot quarrel with Goodlad's (1975) view that the intraschool culture is central to effecting changes in the schools. Apparently inherent in this view is the notion that outside change agents must bring school personnel into some sort of partnership to secure changes whose benefits only the cultivated perceptions of the outsiders are capable of grasping. However, the intraschool culture itself is in the process of change, not in response to the education R&D community per se but rather to larger forces at work in society and the classroom. The trend in the schools is toward exploiting the mastery model for instruction. Progress toward this objective is hampered by a dearth of pertinent tools.

Goodlad and Anderson (1963) envisioned the nongraded school as an alternative to the age/graded school at about the time the age/graded school was sliding into a slow but unremitting decline. Increasingly during the last decade, the schools have responded to legendary deficiencies of the age/grade model by increasing the personalization of instruction--e.g., through team-teaching within a framework of multi-age/grade combinations. The emerging elementary school falls somewhere between the age/graded and the nongraded school--a locus that apparently is apt to transiting students through instruction consonant with the provisions of mastery criteria. NRT adherents have paid little attention to movement of the intraschool culture toward mastery instruction. MAT adherents too often have viewed the trend as the creature of their early efforts and have considered these promising but flawed efforts as sufficiently supportive of the trend.

The schools are in philosophic transition from the age/graded instructional model to a mastery model that bases psychosocial advance on absolute elapsed time (or aging) but instructional advance on a student's proficiency indications. The rate of this desirable transition is hampered by both economic-professional differences and technical insufficiencies. There is a labor-management disagreement that "management" self-servingly expresses in terms of teacher accountability. There is the issue concerning the level in a skills hierarchy at which state agencies should mandate the common instruction. It is possible that the resolution of such issues turns appreciably on development and installation of the technical means required to exploit the mastery model at whatever level is warranted.

Possibly excepting public-industry managers--whose longevity is well-known--professionals on any payroll will be evaluated by one means or another. Mastery instruction inevitably will alter the criteria used to evaluate school personnel at all levels. Children who move from one element of instruction to another in consequence of attaining mastery status for the proficiency taught by the first element lay down clear records of progress. Points in these records can and should be used to determine educational output. In all organizations there are those who would rather be evaluated on the basis of maneuvering and charisma than on a performance basis. Yet, mastery-modelled output has implications that teachers in particular should rather universally favor--e.g.:

- Base lining. Every child delivered to the stewardship of a given teacher--at the moment of delivery or soon thereafter--is characterized for entering achievement in all pertinent skills hierarchies. Thus, all periods of stewardship are marked by base line profiles for prior achievement--particularly referencing the common instruction. The teacher gets credit for all movements beyond base line positions during the period of stewardship. This is much fairer than the wild-card situation that now prevails for entry achievement.

• Fine-tuning. Where the effects of instruction are suitably-often assessed against mastery criteria, the data enable the fine-tuning of instructional management to minimize incipient and unduly delayed decisions to advance students to next instruction. If most features of a system providing instructional status information are automated, teachers should tend to register increases in output that occur without corresponding increases in teacher effort. Instructional effort held constant, speedier and more-explicit feedback for effects probably must increase output.

Such benefits entail technical support for mastery assessment that the schools as yet either do not have or have in insufficient amount. One might attribute letter-grading practices to traditional orientations of teachers and parents. Still, it is evident in schools which extensively team-teach that the best alternative to letter-grading of achievement they can effect as yet is the "vacuum-cleaner nit report"--the bottom line of taxonomic hierarchies, targeted by suppliers of such MATs as yet find their way into the schools.

One glaring deficiency of age/graded instruction is its tendency to repress bad news. Productive education requires both the transmission and the meaningful pinpointing of bad news. The dual achievement-effort reporting system is an appropriate step in the direction of meaningful reporting to parents, since it transfers student accountability to the effort domain. It will continue an undermeaningful response to the progress-reporting requirement until letter grades are replaced by statements of progress made during the reporting period through pertinent skills hierarchies. When this can be done, the letter-grading of effort should be reviewed. A high letter grade for effort referencing a given skills hierarchy asserts the teacher's belief that student participation in instruction was exemplary and that the student could not have completed additional instruction during the reporting period. A lower letter grade for effort must signify a belief that the student could successfully have negotiated X additional subunits of instruction during the reporting period. While the assessment of effort probably cannot be made less intuitive through development and installation of technically superior mastery achievement testing programs, such programs promise a basis for explicating the meaning of such intuitions. When achievement-effort reporting is tied to technically-sufficient assessment of mastery instructional effects in the schools, it should be possible to present bad news to parents in a form they can understand and to use reports containing such news optimally constructively in teacher-parent conferences and other interactions. Noted earlier, most parents are entirely predisposed to advance their children scholastically --particularly regarding the basic proficiencies of common instruction. They cannot be expected to contribute optimally to this objective unless much better informed than they are now concerning the precise domains in which the rate of progress might be improved.

The schools are increasingly predisposed to instruct to mastery criteria. Obtaining the full benefits of this orientation is conditional on bringing into the schools an easily used, cost-effective system for providing instructional status information on a quite-frequent basis. Such a system features inexpensive automating hardware and a criterion-referenced testing program that responds to local perceptions concerning the details of common instruction while exhaustively addressing the pertinent skills hierarchies.

Existing published CRT materials tend to achieve flexibility by treating elemental skills as independent entities and by ignoring assembly skills. The problem is to achieve such flexibility while exhaustively addressing the pertinent skills.

Developing a criterion-referenced testing program appropriate to the responsibilities of a single teacher is a formidable undertaking. Tests developed by teachers on the basis of handbook guidance--e.g., Bloom et al. (1971)--probably should be attempted only as a last resort. Technical assistance of efforts by schools to extend the benefits of mastery instruction probably must be provided both for CRT programs that teachers can adapt to local instructional architectures and for the automating hardware which insures that mastery assessment neither adds clerical burdens to classroom teaching nor new major costs.

The mass market for suitable CRT programs has been developing for a decade. The rate and extent to which it continues to develop --in breadth and depth--now primarily depends on the quality and sensitivity of the technical support for mastery assessment. If such support is responsive and responsible, we might be able largely to dispense with defining strategies for engineering the intraschool culture. I think the larger problem might be to modify the test construction culture--both right and left of center.

References

Bauer, R. A. (Ed.) Social indicators. Cambridge, Mass.: M.I.T. Press, 1966.

Benson, C. S. The economics of public education. Cambridge, Mass.: Riverside Press, 1961.

Bloom, B. S. et al. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Boudon, R. Education, opportunity, and social inequality. New York: Wiley, 1973.

Bowles, S. and Levin, H. M. The determinants of scholastic achievement: An appraisal of some recent evidence. Journal of Human Resources, 1968, 3, 3-24.

Broudy, E. The trouble with textbooks. Teachers College Record, 1975, 77, 13-35.

Bureau of the Census. Statistical abstract of the United States: 1974. Washington, D. C.: Government Printing Office, 1974.

Committee on Assessing the Progress of Education. Citizenship objectives. National Assessment Office, Ann Arbor, Michigan, 1969.

Coleman, J. S. et al. Equality of educational opportunity. Washington, D. C.: Government Printing Office, 1966.

Coleman, J. S. Methods and results in the IEA studies of effects of school on learning. Review of Educational Research, 1975, 45, 355-386.

Comber, L. C. and Keeves, J. P. Science education in nineteen countries. International Studies in Evaluation: Volume 1. Stockholm: Almqvist and Wiksell, 1973.

Gagné, R. M. The conditions of learning. New York: Holt, Rinehart and Winston, 1970 (2nd Edition).

Goodlad, J. I. A perspective for accountability. Phi Delta Kappan, October 1975, 108-112.

Goodlad, J. I. and Anderson, R. H. The nongraded elementary school. New York: Harcourt, Brace and World, 1963 (Revised Edition).

Hanson, R. A. and Schutz, R. E. The effects of programmatic R&D on schooling and the effects of schooling on students: Lessons from the first-year installation of the SWRL/Ginn Kindergarten Program. Technical Report 53, SWRL Educational Research and Development, Los Alamitos, California, 1975.

Harris, C. W. et al. (Eds.) Problems in criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Hively, W. et al. A universe defined system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Hively, W. et al. Domain-referenced curriculum evaluation. Center for the Study of Evaluation, University of California, Los Angeles, 1973.

Jencks, C. et al. Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books, 1972.

Kuhn, T. S. The structure of scientific revolutions. Chicago: University of Chicago Press, 1962.

Maeroff, G. I. H.E.W. Project disputed on testing of students. New York Times, October 5, 1975.

National Assessment of Educational Progress. Mathematics objectives. National Assessment Office, Ann Arbor, Michigan, 1970.

National Center for Educational Statistics. The condition of education. Washington, D. C.: Government Printing Office, 1975.

National Council of Teachers of English. Common sense and testing in English. National Countil of Teachers of English, Urbana, Illinois, 1975.

Nollen, S. D. The economics of education: Research--results and needs. Teachers College Record, 1975, 77, 51-77.

Northcutt, N. Functional literacy in adults: A status report of the Adult Performance Level study. Paper presented to IRA Annual Meeting, New Orleans, May 1-4, 1974.

Northcutt, N. (Final report on APL study, described in Education Daily, October 30, 1975; availability promised in about two months.) 1975 (in press).

Popham, W. J. (Ed.) Criterion-referenced measurement. Englewood Cliffs, N. J.: Educational Technology Publications, 1971.

Rogers, D. C. and Ruchlin, H. S. Economics and education. New York: Free Press, 1971.

Science Research Associates, Inc. Catalog of objectives: SOBAR reading. SRA Criterion-Referenced Measurement Program, Science Research Associates, Inc., Chicago, Illinois, 1974.

Shoemaker, D. M. Principles and procedures of multiple matrix sampling. Cambridge, Massachusetts: Balinger, 1973.

Thomas, J. A. The productive school. New York: Wiley, 1971.

Vandermyn, G. National Assessment achievements: Findings, interpretations, and uses. ECS Report No. 48, Education Commission of the States, Denver, Colorado, 1974.

Wynne, E. Education research: A profession in search of a constituency. Phi Delta Kappan, 1970, 52, 245-247.

Wynne, E. The politics of school accountability. Berkeley, California: McCutchan Publishing Corporation, 1972.

U. S. Department of Health, Education, and Welfare. Toward a social report. Washington, D. C.: Government Printing Office, 1973.